Using Aggregate Geographic Data to Proxy Individual Socioeconomic Status: Does Size Matter?

Mah-jabeen Soobader, PhD, Felicia B. LeClere, PhD, Wilbur Hadden, PhD, and Brooke Maury, BA

A B S T R A C T

Objectives. This study assessed whether aggregate-level measures of socioeconomic status (SES) are less biased as proxies for individual-level measures if the unit of geographic aggregation is small in size and population.

Methods. National Health Interview Survey and census data were used to replicate analyses that identified the degree to which aggregate proxies of individual SES bias interpretations of the effects of SES on health.

Results. Ordinary least squares regressions on self-perceived health showed that the coefficients for income and education measured at the tract and block group levels were larger than those at the individual level but smaller than those estimated by Geronimus et al. at the zip code level.

Conclusions. Researchers should be cautious about use of proxy measurement of individual SES even if proxies are calculated from small geographic units. (*Am J Public Health.* 2001;91:632–636)

Aggregate proxies have often been used for uncollected or missing information on an individual's socioeconomic status (SES) in analyses of health outcomes.[1–6] In related analyses, however, Geronimus and Bound[7] and Geronimus et al.[8] have demonstrated that these geographic or aggregate proxies contain 2 types of statistical bias—errors-in-variables bias and aggregation bias—that affect estimates of the impact of SES on health and may confound the effects of other predictors correlated with SES. In a comparison of zip code and census tract data, Geronimus and Bound observed that measures from smaller geographic units may introduce smaller biases, although the gains could be limited. In other research, there has been considerable ambivalence as to whether gains can be made from using smaller aggregates to generate appropriate proxies for individual SES.[1,4,7,9–11]

Logically, estimates from smaller, more homogeneous geographic units should provide closer approximates of individual SES, but there is no clear consensus. Census tracts are not only a smaller but also a more homogeneous aggregate unit than zip codes, which are assigned by the US Postal Service merely for convenience and can cross county and state boundaries. By contrast, census tracts are designed to be homogeneous with respect to population characteristics, economic status, and living conditions. Yet, the gains from focusing on tracts instead of zip codes appear small.

In this report, we extend the analyses of Geronimus and colleagues to more fully examine the assumption that estimates from smaller geographic aggregates yield better SES proxies for individuals and to assess the consequences of such an assumption. We used the methods outlined by Geronimus et al. to extend the analytic question asked by Geronimus and Bound: Does the size of the aggregate proxy influence the amount of bias introduced? In addition, we addressed recent criticisms of the study of Geronimus and Bound suggesting that the data sets they used were nonrepresentative, poor in regard to geographic coding ("geocoding"), and inadequate in size.[2,12]

Our replication involved a large, nationally representative survey in which about 90% of respondents were geocoded and the response rate was approximately 95%. The data were derived from 3 years (1988–1990) of the National Health Interview Survey (NHIS) linked to data from the 1990 census at the census tract and block group levels. In the 1990 census, census tract and block group enumeration was extended to all US counties through the block numbering area system.[13] Analyses based on previous censuses, including that of Geronimus and Bound, did not include rural areas, because before 1990 these areas were largely excluded from the tracting process.

Block groups tend to be more heterogeneous in nonmetropolitan areas, and our ability to include areas outside of cities improved the generalizability of the results. We reproduced the analyses of Geronimus et al. for tracts and block groups in the continental United States and compared the size of the coefficients and the direction of the biases inherent in predictions of self-perceived health.

Methods

In this study, data from the NHIS were concatenated for 3 consecutive years, 1988 through 1990. The NHIS is an annual survey of the civilian noninstitutionalized population of the United States using a stratified multistage probability sampling method. Detailed information about the demographic and socioeconomic characteristics and health status of a nationally representative sample of individuals is collected.[14] These data are supplemented with data derived from census files containing economic and demographic information on states and geographic subdivisions. The 2 files are linked by geocodes, which are attached to individual NHIS records.

The sample consisted of all non-Hispanic respondents 18 years or older. All individual measures were derived from the NHIS. Re-

Mah-jabeen Soobader is with the Department of Pediatrics, University of Rochester, Rochester, NY. Felicia B. LeClere is with the Laboratory for Social Research and Department of Sociology, University of Notre Dame, Notre Dame, Ind. Wilbur Hadden and Brooke Maury are with the Office of Program Development and Extramural Programs, National Center for Health Statistics, Hyattsville, Md.

Requests for reprints should be sent to Felicia B. LeClere, PhD, Laboratory for Social Research and Department of Sociology, University of Notre Dame, 913 Flanner Hall, Notre Dame, IN 46556 (e-mail: leclere.1@nd.edu).

This brief was accepted July 18, 2000.

spondents were classified as non-Hispanic Black or non-Hispanic White. For comparative purposes, we used the 5-point self-perceived health measure used by Geronimus et al. in its continuous form (1 referring to poor health and 5 referring to excellent health). To provide an accurate replication of the original study of Geronimus and colleagues,[8] we assumed that "ordered categorical responses reflect an underlying latent continuous variable."[8(p531)] Individual SES measures included continuous years of completed schooling and log of income-to-needs ratio. The midpoints of the income categories used in the NHIS, which are adjusted for family size, were used in calculating income-to-needs ratios. This measure provided comparability across households of different sizes. Controls for sex and age were also included in the multivariate analyses. The concatenated sample included 183 706 individuals, among whom 12.2% reported being in fair or poor health and 34.9% reported being in excellent health.

Two variables derived from the 1990 census were used at the tract and block group levels: log of median household income of the tract or block group and mean educational attainment of the inhabitants of the geographic aggregate. To calculate mean education levels, we used grouped census data and estimated means by multiplying the number of individuals in each category by the midpoint of the category. These values were summed over the categories for estimates of total number of years of education, which was subsequently divided by the number of persons in the geographic unit to obtain the mean educational attainment in that unit. Data from 11 477 block groups and 8667 tracts were used in this analysis.

The analysis was divided into 2 parts. The first part involved estimation of ordinary least squares regressions for self-perceived health from the individual and aggregate proxies for educational attainment and income. The purpose of these regressions was 3-fold. First, the use of 3 measurement strategies (individual, tract, and block group) in comparisons between the coefficients for education and income across the equations estimated helps illustrate the size and direction of the bias in each of the 2 aggregate measures. Second, differences in the coefficient attached to race demonstrate the ability of different types of SES measures to control for SES confounding in the relationship between health and race. Finally, differences in the overall fit of the model demonstrate the explanatory strength of different measures. This strategy is consistent with the method of interpreting results used by Geronimus and colleagues.

The second part of the analysis was designed to assess the degree, direction, and nature of the bias in the aggregate proxies. We replicated the Geronimus et al. decomposition in 2 steps. First, we regressed individual income and education on the aggregate proxies as well as on race, age, and sex, which provided empirical estimates of errors-in-variables bias. Second, we regressed the residuals from the health regressions (with individual SES measures estimated in the first part of the analysis) on the aggregate proxies and other covariates at both the tract and block group levels, which provided an estimate of the magnitude and direction of the coefficient bias due to aggregation. We compared these estimated biases with those reported by Geronimus et al. for zip codes. Although we used different samples from different years, our results were substantially similar, and thus we believe that the comparisons are warranted.

Results

Means, standard deviations, and Pearson correlation coefficients for the individual and aggregate proxies appear in Table 1. These descriptive results support Geronimus and Bound's conclusions concerning aggregate proxies. Aggregate income and education means were higher and showed less variability than similar measures at the individual level. There was only a slight difference between tract and block group estimates, with no significant trend by size of aggregate unit.

Individual SES and aggregate proxy measures were moderately correlated (0.33–0.44) and of similar size across tract and block groups. Income and education were highly correlated (0.69–0.92) within and across the geographic aggregates but only moderately correlated at the individual level (0.39). The consistency in the direction and size of the sample statistics suggests that, at least at the descriptive level, the size of the geographic aggregate did not significantly improve the accuracy of the SES proxy measure.

The results of the ordinary least squares regressions for self-perceived health appear in Table 2. The first column contains the coefficient for race and controls for age and sex. The next 3 columns add controls for income, education, and the combination of income and education using (successively) individual, tract, and block group measures.

The coefficients attached to the SES proxies in regressions in which they were entered singly were consistently higher than those attached to the individual measures. The income coefficient was between 33% and 50% larger, and the education coefficient was between 46% and 52% larger. The block group coefficients were consistently smaller than those estimated

TABLE 1—Sample Statistics: National Health Interview Survey, 1988–1990

			Pearson Correlation Coefficients							
			Individual				Aggregate: Tract		Aggregate: Block Group	
Variable Level	Mean	SD	Health	Race (White)	Log Family Income	Education	Log Median Income	Mean Education	Log Median Income	Mean Education
Individual										
Health	3.83	1.11	1.00							
Race (White)	0.86	0.35	0.13	1.00						
Log family income	9.24	0.85	0.24	0.27	1.00					
Education	12.63	2.85	0.34	0.15	0.39	1.00				
Aggregate										
Tract										
Log median income	10.45	0.41	0.21	0.38	0.43	0.33	1.00			
Mean education	12.97	1.41	0.22	0.27	0.37	0.40	0.77	1.00		
Block group										
Log median income	10.44	0.46	.21	0.37	0.44	0.33	0.89	0.69	1.00	
Mean education	12.97	1.25	.22	0.26	0.38	0.41	0.71	0.92	0.72	1.00

TABLE 2—Ordinary Least Squares Estimates of the Effects of Race and Socioeconomic Status (SES) Proxies on Self-Perceived Health

Explanatory Variable	SES Proxy, Coefficient (SE)									
		Individual[a]			Tract[a]			Block Group[a]		
	None	Income	Education	Income + Education	Income	Education	Income + Education	Income	Education	Income + Education
White	0.422 (0.007)	0.241 (0.007)	0.314 (0.007)	0.209 (0.007)	0.233 (0.007)	0.292 (0.007)	0.243 (0.007)	0.234 (0.007)	0.291 (0.007)	0.235 (0.007)
Income		0.318 (0.003)		0.220 (0.003)	0.476 (0.006)		0.232 (0.009)	0.431 (0.006)		0.222 (0.008)
Education			0.100 (0.001)	0.074 (0.001)		0.172 (0.002)	0.112 (0.003)		0.160 (0.002)	0.105 (0.003)
R^2	0.128	0.183	0.189	0.211	0.154	0.157	0.160	0.211	0.155	0.158

Note. Data were derived from the National Health Interview Survey, 1988–1990, and 1990 census files. All equations include controls for sex and age.
[a]The 3 columns below represent coefficients with controls for income, education, and their combination.

for the tracts but remained closer in magnitude to the tracts than to the individual measures. These findings were consistent with both analyses conducted by Geronimus and colleagues. Only a slight advantage was gained in smaller aggregate units, while the size of the difference remained relatively large between aggregate and individual measures. Individual model R^2 values were slightly higher, while the models with aggregate measures did not fit as well, with the exception of the median income block group model.

Our analysis of the confounding of race with SES was not consistent with previous analyses, however. Geronimus et al. suggested that aggregate proxies do not sufficiently control for SES confounding in the interpretation of race effects on health, which may cause analysts to overstate the independent effect of race. They supported this conclusion by demonstrating that the size of the coefficient attached to race is not reduced by the same degree with the introduction of aggregate SES proxies as it is with individual measures.

In our analysis, the tract and block group measures were as effective as the individual measures in controlling for SES confounding in the interpretation of the effect of race on health outcomes, with the exception of the final equation in each set, which included education and income simultaneously. In the case of tracts and block groups, median aggregate income alone did a better job of accounting for the effect of race than did the combination. One method for reconciling this result with the findings of Geronimus and colleagues involves differences in the racial and economic homogeneity of the aggregate unit. The advantage of smaller, consistently defined aggregate units may be their degree of demographic homogeneity, which increases the likelihood that the aggregate proxy captures both individual effects and "contextual" or neighborhood effects of SES.

Data on the decomposition of the source and direction of the coefficient bias (as described by Geronimus et al.) appear in Table 3. The top section contains the decomposition for the tract-level measures, and the bottom section includes data on the block group measures. The first 2 sets of equations in each subsection contain the coefficients for the regressions of race and the aggregate SES proxy singly or in combination after control for age and sex. These equations can be used to assess the degree to which the lack of a perfect correlation between aggregate and individual measures of SES affects the predicted impact of SES on health and fails to alleviate the confounding of race with income (errors in variables).

The final column in each subsection of Table 3 contains the regressions for the residuals from the equations that appear in Table 2; these regressions include controls for individual SES and other demographic characteristics. These values provide an indication of the

TABLE 3—Bias Decomposition for Aggregate Socioeconomic Status (SES) Proxies: Census Tract and Block Group

Explanatory Variable	Income,[a] Coefficient (SE)			Education,[b] Coefficient (SE)			Income and Education, Coefficient (SE)		
	Individual Income		Residual	Individual Education		Residual	Income	Education	Residual
SES proxy: tract									
Race (White)	0.631 (0.005)	0.275 (0.005)	–0.095 (0.007)	1.285 (0.018)	0.465 (0.018)	–0.067 (0.007)	0.281 (0.005)	0.418 (0.018)	–0.059 (0.007)
Aggregate proxy									
Income		0.812 (0.005)	0.218 (0.006)				0.647 (0.007)	0.220 (0.024)	0.074 (0.009)
Education					0.941 (0.005)	0.078 (0.002)	0.076 (0.002)	0.884 (0.008)	0.030 (0.003)
R^2	0.082	0.212	0.007	0.084	0.215	0.007	0.216	0.215	0.003
SES proxy: block group									
Race (White)	0.631 (0.005)	0.272 (0.005)	–0.094 (0.007)	1.285 (0.018)	0.450 (0.017)	–0.067 (0.007)	0.273 (0.005)	0.351 (0.018)	–0.060 (0.007)
Aggregate proxy									
Income		0.741 (0.004)	0.195 (0.006)				0.565 (0.006)	0.393 (0.019)	0.069 (0.008)
Education					0.883 (0.005)	0.072 (0.002)	0.089 (0.002)	0.786 (0.007)	0.028 (0.003)
R^2	0.082	0.218	0.007	0.084	0.223	0.008	0.226	0.225	0.003

[a]The first 2 columns regress log of family income on race and aggregate proxies, respectively. The third column regresses the residuals from the individual-level equations in Table 2 on race and aggregate proxies.
[b]The first 2 columns regress educational attainment on race and aggregate proxies, respectively. The third column regresses the residuals from the individual-level equations in Table 2 on race and aggregate proxies.

degree to which proxy measures of SES assess a larger concept than individual SES measures (aggregation bias).

Our results suggest that the smaller aggregates, with 2 exceptions, systematically reduced the degree of empirical bias introduced by the use of aggregate proxies for individual SES in individual-level analyses. However, all sources of bias remained significant and substantial, even for measures based on the block group aggregate, which contained an average of approximately 1000 people. Thus, although there may have been gains from using smaller geographic areas, the bias remained.

The first issue identified by Geronimus and colleagues was the extent to which the estimate of the effects of income (education) on health may be affected by the use of a proxy. The coefficients attached to income in columns 2 and 5 of Table 3 indicate the degree to which the estimated coefficient may have been downwardly biased owing to the limited correlation between the individual and aggregate proxy (errors in variables). The more removed the coefficients in columns 2 and 5 from 1, the larger the downward bias in the estimated relationship between the SES indicators and health (see Geronimus et al.[8] for a full description of the calculation of these effects).

In the case of income, the tract (0.812) and block (0.741) proxies actually introduced more downward bias than the zip code, as estimated by Geronimus and colleagues (0.983). In the case of education, the tract (0.941) and the block group (0.883) both introduced less bias than the zip code (0.779). While logically the result for income seems counterintuitive, it may be that larger aggregate units provide more robust estimates because the sample is larger, which improves the correlation between individuals and aggregates. Also, the aggregate proxy may measure a broader concept of economic status than individual SES (aggregation bias), and thus the aggregate proxy may have a larger impact on health than would be expected for the individual SES measure. The coefficients attached to income and education in the third and sixth columns of Table 3 are estimates of this upward bias. The larger the positive estimate (given an additive rather than a multiplicative effect), the more bias introduced. In this case, smaller aggregates unambiguously introduced less bias for income (0.318 for zip codes, 0.218 for tracts, and 0.195 for block groups) but not for education (0.068, 0.078, and 0.072, respectively).

The second issue identified by Geronimus and colleagues was whether the use of aggregate proxies imperfectly controls the confounding of race with income in analyses of the impact of race on health. Errors-in-variables bias (measured by the race coefficients in columns 2 and 5 of Table 3) scales the amount of bias introduced by the correlation between race and income. The smaller the coefficient, the less bias introduced by the proxy measure. In the case of both income (0.412 for zip code [as reported by Geronimus et al.], 0.275 for tract, and 0.272 for block group) and education (0.522, 0.465, and 0.450, respectively), the smaller aggregates introduced less bias.

Block groups and tracts also reduced the amount of SES confounding with race because they may have captured a more inclusive measure of economic status than individual SES, and it is likely that they were the appropriate aggregate unit in which to assess this measure. The coefficients attached to race in the third and sixth columns of Table 3 describe the degree to which confounding may have been reduced by the addition of the aggregate measure. The more negative the coefficient, the larger the reduction in bias. Again, the smaller aggregates had the advantage of reducing the bias more than the larger aggregates for both income (–0.086 for zip codes, –0.095 for tracts, and –0.094 for block groups in Geronimus et al.[8]) and education (0.068, 0.078, and 0.072, respectively), although the relationship was not monotonic.

Discussion

This replication emphasizes 2 main points that will be of use to researchers who are contemplating using geographic proxies for individual measures of SES. First, these findings support the original work of Geronimus and colleagues, which demonstrated that aggregate proxies cannot be used without acknowledgment of the potential bias introduced in estimating coefficients attached to SES measures. This replication, which provides surprisingly consistent results given the differences in data collection, does not suffer from any of the potential limitations identified by critics of Geronimus and Bound (e.g., selective loss to follow-up, limited geocoding).

Second, this research suggests that although the magnitude of the bias is substantial, it decreases with the size of the geographic unit for most sources of bias. Therefore, although researchers using aggregate proxies clearly cannot ignore the inherent bias, measures from smaller geographic units may produce results for aggregate SES measures that are slightly less biased.

This analysis does not suggest that individual measures of SES alone are perfect measures of economic status or access to resources that produce good health. We found, conversely, that in some cases the aggregate proxy is more likely to control SES confounding in the relationship between SES and race. This is consistent with emerging research on the effect of neighborhoods on the health of minorities.[15–18] Thus, researchers should also be cautious in their use of individual measures alone. □

Contributors

M. Soobader contributed to conceptualization of the problem, analysis of the data, and the writing of the manuscript. F. B. LeClere contributed to conceptualization of the problem and the writing of the manuscript. W. Hadden and B. Maury contributed to conceptualization of the problem and data analysis.

References

1. Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *Am J Public Health.* 1992;82:703–710.
2. Krieger N, Gordon D. Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples [letter]. *Am J Epidemiol*. 1999;148:475–486.
3. Krieger N. Social class and the black/white crossover in the age-specific incidence of breast cancer: a study linking census-derived data to population-based registry records. *Am J Epidemiol*. 1990;131:804–814.
4. Krieger N. Women and social class: a methodological study comparing individual, household, and census measures as predictors of black/white differences in reproductive history. *J Epidemiol Community Health*. 1991;45:35–42.
5. Krieger N, Williams DR, Moss NE. Measuring social class in US public health research. *Annu Rev Public Health*. 1997;18:341–378.
6. Moss N, Krieger N. Measuring social inequalities in health. *Public Health Rep.* 1995;110:302–305.
7. Geronimus AT, Bound J. Use of census-based aggregate variables to proxy for socioeconomic group: evidence from national samples. *Am J Epidemiol.* 1998;148:475–486.
8. Geronimus AT, Bound J, Neidert LJ. On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. *J Am Stat Assoc.* 1996;91:529–537.
9. Danesh J, Gault S, Semmence J, Appleby P, Peto R. Postcodes as useful markers of social class: population based study in 26,000 British households. *BMJ.* 1999;318:843–845.
10. Greenwald HP, Polissar NL, Borgatta EF, McCorkle R. Detecting survival effects of socioeconomic status: problems in the use of aggregate measures. *J Clin Epidemiol.* 1994;47:903–909.
11. Hyndman JC, Holman CD, Hockey RL, Donovan RJ, Corti B, Rivera J. Misclassification of social disadvantage based on geographical areas: comparison of postcode and collector's district analyses. *Int J Epidemiol.* 1995;24:165–176.
12. Davey Smith G, Hart C, Ben-Shlomo Y. Use of census-based aggregate variables to proxy socioeconomic group: evidence from national

samples [letter]. *Am J Epidemiol.* 1999;148: 996–997.
13. *Geographic Areas Reference Manual*. Washington, DC: US Bureau of the Census; 1994.
14. Adams PF, Benson V. Current estimates from the National Health Interview Survey. *Vital Health Stat 10.* 1991;No. 181.
15. Anderson RT, Sorlie P, Backlund E, Johnson N, Kaplan GA. Mortality effects of community socioeconomic status. *Epidemiology*. 1997;8: 42–47.
16. LeClere FB, Rogers RG, Peters KD. Ethnicity and mortality in the United States: individual and community correlates. *Soc Forces.* 1997;76:169–198.
17. LeClere FB, Rogers RG, Peters K. Neighborhood social context and racial differences in women's heart disease mortality. *J Health Soc Behav.* 1998;39:91–107.
18. Waitzman NJ, Smith KR. Phantom of the area: poverty-area residence and mortality in the United States [published correction appears in *Am J Public Health.* 1998;88:1122]. *Am J Public Health*. 1998;88:973–976.